# Generalization in Federated Learning: A Conditional Mutual Information Framework

Ziqiao Wang [1] Cheng Long [2] Yongyi Mao [3]

[1]Tongji University [2]Columbia University [3]University of Ottawa

## Two-Level Generalization in Federated Learning (FL)

- $K$ participating clients $\{\mu_i\}_{i=1}^K \sim \mathcal{D}$
- Each local learning algorithm $\mathcal{A}_i : \mathcal{Z}^n \to \mathcal{W}$ i.e. mapping $S_i = \{Z_{i,j}\}_{j=1}^n \sim \mu_i$ to a local model $W_i$.
- Central server: merging $\{W_i\}_{i=1}^K$ to obtain $W$
- Generalization Error:

$$\mathcal{E}_{\mathcal{D}}(\mathcal{A}) = \underbrace{\mathbb{E}_W\left[L_{\mathcal{D}}(W)\right] - \mathbb{E}_{W,\mu_{[K]}}\left[L_{\mu_{[K]}}(W)\right]}_{\mathcal{E}_{PG}(\mathcal{A})\text{:Participation Gap}} + \underbrace{\mathbb{E}_{W,\mu_{[K]}}\left[L_{\mu_{[K]}}(W)\right] - \mathbb{E}_{W,S}\left[L_S(W)\right]}_{\mathcal{E}_{OG}(\mathcal{A})\text{:Out-of-Sample Gap}}.$$

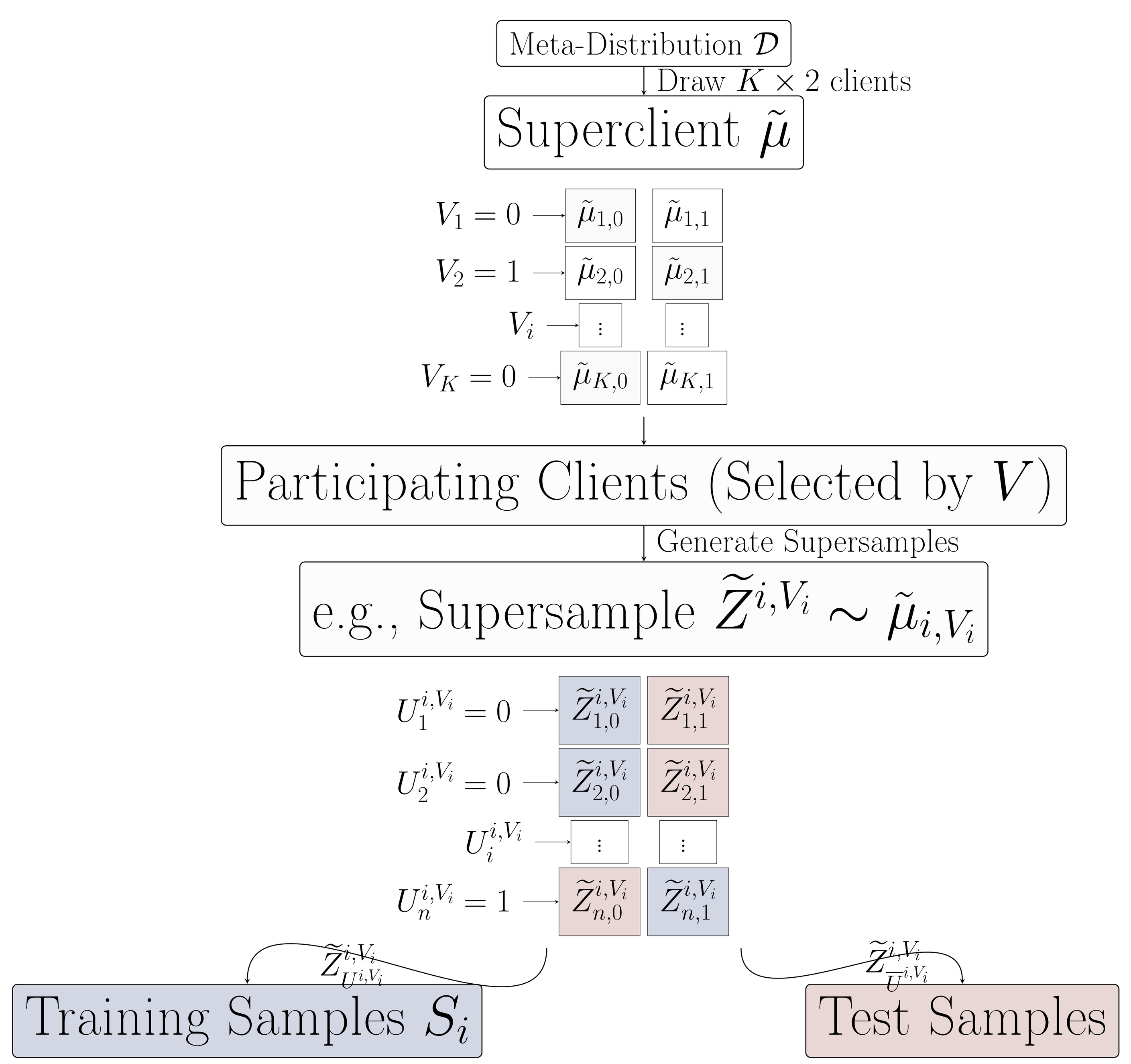
$L_{\mathcal{D}}(w) \triangleq \mathbb{E}_{\mu\sim\mathcal{D}}\mathbb{E}_{Z\sim\mu}[\ell(w,Z)]$: global true risk

$L_{\mu_{[K]}}(w) \triangleq \frac{1}{K}\sum_{i=1}^K \mathbb{E}_{Z_i'\sim\mu_i}[\ell(w,Z_i')]$: average client true risk

$L_S(w) \triangleq \frac{1}{Kn}\sum_{i=1}^K\sum_{j=1}^n \ell(w,Z_{i,j})$: average client empirical risk

## Superclient and Supersample Construction in FL

Similar to [4], we construct superclient $\tilde{\mu}$ and supersamples $\widetilde{Z}$:




## Main Contributions

- We derive the first CMI-based bound for FL, consisting of two terms: (i) the CMI between the hypothesis and the Bernoulli variable governing client participation, and (ii) the CMI between the hypothesis and the Bernoulli variable governing training data membership.
- We derive evaluated CMI (e-CMI) bounds, which recover the best-known FL convergence rate in the low empirical risk regime, and are easy to measure.
- For model averaging and structured loss functions, we obtain fast-rate convergence w.r.t. the number of participating clients

## Conditional Mutual Information Bounds for FL

### Lemma 1 (informal): Symmetric Properties

$$\mathcal{E}_{PG}(\mathcal{A}) = \frac{1}{K}\sum_{i=1}^K \mathbb{E}\left[(-1)^{V_i}\left(\ell(W,\widetilde{Z}^{i,1}_{1,\overline{U}_1^{i,1}}) - \ell(W,\widetilde{Z}^{i,0}_{1,\overline{U}_1^{i,0}})\right)\right]$$

$$\mathcal{E}_{OG}(\mathcal{A}) = \frac{1}{Kn}\sum_{i=1}^K\sum_{j=1}^n \mathbb{E}\left[(-1)^{U_j^{i,V_i}}\left(\ell(W,\widetilde{Z}^{i,V_i}_{j,1}) - \ell(W,\widetilde{Z}^{i,V_i}_{j,0})\right)\right]$$

### First CMI Bound

- Assume the loss $\in [0,1]$

$$\boxed{|\mathcal{E}_{\mathcal{D}}(\mathcal{A})| \leq \sqrt{\frac{2I(W;V|\widetilde{Z},U)}{K}} + \sqrt{\frac{2I(W;U|\widetilde{Z},V)}{Kn}}.}$$

$$\Downarrow$$

- Differential Privacy (DP) & Generalization

$$\begin{cases} \frac{I(W;V|\widetilde{Z},U)}{K} \leq \frac{\min\{\epsilon',(e^{\epsilon'}-1)\epsilon'\}}{K} & \text{if the overall algorithm is } \epsilon'\text{-DP,} \\ \frac{I(W;U^{i,V_i}|\widetilde{Z}^i,V_i)}{n} \leq \frac{\min\{\epsilon_i,(e^{\epsilon_i}-1)\epsilon_i\}}{n} & \text{if } \mathcal{A}_i \text{ is } \epsilon_i\text{-DP.} \end{cases}$$

- If all clients have the same $\mu$, we have $|\mathcal{E}_{\mathcal{D}}(\mathcal{A})| \leq \sqrt{\frac{2I(W;U|\widetilde{Z})}{Kn}}$.

### High Probability CMI Bound

W.p. $\geq 1-\delta$ under the draw of $(\widetilde{Z},U,V)$, gen. error of FL is upper bounded by

$$\sqrt{\frac{\mathrm{D_{KL}}\left(P_{W|\widetilde{Z},U,V}||P_{W|\widetilde{Z},U}\right) + \log\frac{\sqrt{K}}{\delta}}{K-1}} + \sqrt{\frac{\mathrm{D_{KL}}\left(P_{W|\widetilde{Z},U,V}||P_{W|\widetilde{Z},V}\right) + \log\frac{\sqrt{Kn}}{\delta}}{Kn-1}}.$$

### Fast-rate CMI BoundS

There exist constants

$C_1,C_2,C_3,C_4 \in \{C_1,C_2 > 1, C_3,C_4 > 0 | e^{-2C_3C_1} + e^{2C_3} \leq 2, e^{-2C_4C_2} + e^{2C_4} \leq 2\}$

s.t.

$$\boxed{\mathbb{E}_W\left[L_{\mathcal{D}}(W)\right] \leq C_1C_2\mathbb{E}_{W,S}\left[L_S(W)\right] + \frac{I(W;V|\widetilde{Z},U)}{C_3K} + \frac{C_1I(W;U|\widetilde{Z},V)}{C_4Kn}.}$$

Following [5], let $\begin{cases} \bar{L}_i^+ = \ell(W,\widetilde{Z}^{i,0}_{1,\overline{U}_1^{i,0}}) \\ L_j^{i+} = \ell(W,\widetilde{Z}^{i,V_i}_{j,0}) \end{cases}$, we also have an e-CMI bound

$$\mathbb{E}_W\left[L_{\mathcal{D}}(W)\right] \leq C_1C_2\mathbb{E}_{W,S}\left[L_S(W)\right] + \sum_{i=1}^K\frac{I(\bar{L}_i^+;V_i)}{C_3K} + \sum_{i=1}^K\sum_{j=1}^n\frac{C_1I(L_j^{i+};U_j^{i,V_i}|V_i)}{C_4Kn}.$$

- If $\mathbb{E}_{W,S}\left[L_S(W)\right]$ is small, then achieving fast rate $\mathcal{O}\left(\frac{1}{K}+\frac{1}{Kn}\right)$
- e-CMI is easy to estimate in practice (e.g., MI between one-dimensional R.V.'s)

## Extension: CMI Bounds for Model Aggregation in FL

- Following FedAvg algorithm [3], the aggregation is $W = \frac{1}{K}\sum_{i=1}^K W_i$.

### Bregman Divergence Loss

- Similar to [1], consider a Bregman divergence loss: $D_f(x,y) \triangleq f(x) - f(y) - \langle\nabla f(y), x-y\rangle$ for a strictly convex function $f$.
- Let $\ell(w,z) = D_f(w,z)$, under some sub-Gaussian conditions,

$$\boxed{|\mathcal{E}_{\mathcal{D}}(\mathcal{A})| \lesssim \frac{1}{K^2}\sum_{i=1}^K \mathbb{E}_{\widetilde{Z}^i,U^i}\sqrt{I^{\widetilde{Z}^i,U^i}(W_i;V_i)} + \frac{1}{K^2n}\sum_{i=1}^K\sum_{j=1}^n \mathbb{E}_{\widetilde{Z}^i,V_i}\sqrt{I^{\widetilde{Z}^i,V_i}(W_i;U_j^{i,V_i})}.}$$

- Now based on local CMI terms (i.e., CMI based on $W_i$)!
- If each client $i$ can only transmit $B$ bits of information to the central server, then

$$|\mathcal{E}_{\mathcal{D}}(\mathcal{A})| \leq \mathcal{O}\left(\frac{\sqrt{B}}{K} + \frac{1}{K}\sqrt{\frac{B}{n}}\right).$$

$\Longrightarrow$ Fast-rate behavior w.r.t. the number of clients.

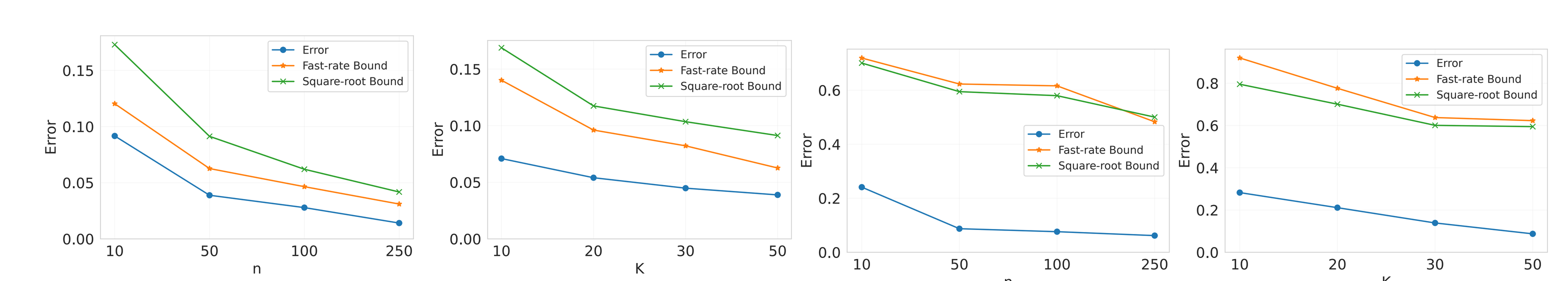
### Smooth and Strongly Convex Loss

- Similar to [2], let $\ell$ be smooth and strongly convex, and let $\xi_{i,j} = \mathbb{E}\left[\ell(W,Z_{i,j})\right]$. If $\mathcal{A}_i$ is an interpolating algorithm (i.e. achieving zero local training loss),

$$|\mathcal{E}_{OG}(\mathcal{A})| \leq \frac{1}{K^3n}\sum_{i=1}^K\sum_{j=1}^n I(W_i;U_j^{i,V_i}|\widetilde{Z}^i,V_i) + \frac{1}{K^2n}\sum_{i=1}^K\sum_{j=1}^n\sqrt{\xi_{i,j}I(W_i;U_j^{i,V_i}|\widetilde{Z}^i,V_i)}.$$

- (i) Even faster decay rate for $|\mathcal{E}_{OG}(\mathcal{A})|$. (ii) Data heterogeneity captured by $\xi_{i,j}$.

## Empirical Results



(a) MNIST ($n$) (b) MNIST ($K$) (c) CIFAR-10 ($n$) (d) CIFAR-10 ($K$)

Figure 1. Verification of bounds on MNIST and CIFAR-10.

## References

[1] Leighton Pate Barnes, Alex Dytso, and H Vincent Poor. Improved information theoretic generalization bounds for distributed and federated learning. In 2022 IEEE International Symposium on Information Theory (ISIT), pages 1465–1470. IEEE, 2022.

[2] Peyman Gholami and Hulya Seferoglu. Improved generalization bounds for communication efficient federated learning. arXiv preprint arXiv:2404.11754, 2024.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017.

[4] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In Conference on Learning Theory. PMLR, 2020.

[5] Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. In International Conference on Machine Learning. PMLR, 2023.